SEII

SCHOOL EFFECTIVENESS &
INEQUALITY INITIATIVE

Discussion Paper #2014.01

Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start

Christopher Walters

April 2014

MIT Department of Economics
77 Massachusetts Avenue, Bldg E19-750
Cambridge, MA 02139

National Bureau of Economic Research
1050 Massachusetts Avenue, 3rd Floor
Cambridge, MA 02138

Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start

Christopher Walters*

School Effectiveness and Inequality Initiative Discussion Paper #2014.01
April 2014

Abstract

Studies of small-scale "model" early-childhood programs show that high-quality preschool can have transformative effects on human capital and economic outcomes. Evidence on the Head Start program is more mixed. Inputs and practices vary widely across Head Start centers, however, and little is known about variation in effectiveness within Head Start. This paper uses data from a multi-site randomized evaluation to quantify and explain variation in effectiveness across Head Start childcare centers. I answer two questions: (1) How much do short-run effects vary across Head Start centers? and (2) To what extent do inputs and practices explain this variation? To answer the first question, I develop a random coefficients sample selection model that quantifies heterogeneity in Head Start effects, accounting for non-compliance with experimental assignments. Empirical Bayes estimates of the model show that the cross-center standard deviation of cognitive effects is 0.18 test score standard deviations, which is larger than typical estimates of variation in teacher or school effectiveness. Next, I assess the role of inputs in generating this variation, focusing on inputs commonly cited as central to the success of model programs. My results show that Head Start centers offering full-day service boost cognitive skills more than other centers, while Head Start centers offering frequent home visiting are especially effective at raising non-cognitive skills. Other key inputs, including the High/Scope curriculum, teacher education and certification, and class size, are not associated with increased effectiveness in Head Start. Together, observed inputs explain a small share of the variation in Head Start effectiveness. An investigation of the role of counterfactual childcare choices suggests that cross-center differences in effects may be partially due to differences in rates of private preschool attendance.

*University of California, Berkeley Economics (e-mail: crwalters@econ.berkeley.edu). I am grateful to Joshua Angrist, Aviva Aron-Dine, David Autor, David Chan, Hilary Hoynes, Guido Imbens, Patrick Kline, Enrico Moretti, Christopher Palmer, Parag Pathak, Jesse Rothstein, Tyler Williams, and participants at the MIT labor economics lunch for useful comments and suggestions. This work was supported by Institute for Education Sciences award number R305A120269 and a National Academy of Education/Spencer Dissertation Fellowship.

1 Introduction

Studies of small-scale "model" early-childhood education programs show that preschool attendance can boost outcomes in the short- and long-run. In the High/Scope Perry Preschool Project, a randomized trial that took place in the early 1960s, 123 disadvantaged children were randomly assigned to either an intensive preschool program or a control group without access to the program. Subsequent analyses showed that participation in the Perry program increased average IQ at age 5 by nearly a full standard deviation, and had lasting impacts on educational attainment, criminal behavior, drug use, employment, and earnings (Anderson 2008; Berruta-Clement et al. 1984; Heckman et al. 2010b; Schweinhart et al. 1997, 2005).[1] Heckman et al. (2010a) estimate the annual social rate of return to the Perry Project at between 7 and 10 percent. The North Carolina Abecedarian Project, another small-scale intervention, had similarly dramatic effects (Campbell and Ramey 1994, 1995). The striking success of these programs has led some analysts to argue that the returns to educational intervention peak early in life (Heckman 2011). These findings have also motivated recent calls for expansion of publicly-provided preschool (Obama 2013).

In contrast, evidence on the effects of large-scale early childhood programs is more mixed. Early quasi-experimental studies of Head Start, the largest early childhood program in the United States, showed positive effects on cognitive skills, child mortality, and long-term outcomes (Currie and Thomas 1995; Ludwig and Miller 2007; Garces et al. 2002; Deming 2009).[2] More recently, results from the Head Start Impact Study (HSIS), the first randomized evaluation of Head Start, showed smaller, less-persistent gains. The HSIS experiment involved random assignment of more than 4,000 children to Head Start or a control group at over 300 childcare centers throughout the US. The HSIS treatment group outscored the control group by roughly 0.1 standard deviations on measures of cognitive skill during preschool, but these gains did not persist into kindergarten (US Department of Health and Human Services 2010, 2012; Bitler et al. 2012). Moreover, the HSIS experiment showed little evidence of effects for a wide range of non-cognitive and health outcomes (US Department of Health and Human Services 2010).[3]

Inputs and practices vary widely across Head Start centers, however, and little is known about variation in effectiveness within Head Start. This paper uses HSIS data to quantify and explain variation in effectiveness across Head Start childcare centers. I focus on one candidate explanation for differences in the effects of early childhood programs: differences in inputs and practices used by these programs. Some Head Start centers use inputs more similar to successful model programs than others. For example, one-third of Head Start centers use the High/Scope curriculum, the centerpiece of the Perry Preschool experiment. Head Start centers also differ with respect to teacher characteristics, class size, instructional time, and frequency of

[1] Anderson (2008) argues that the Perry Project produced significant long-term benefits only for girls.

[2] Other studies finding positive effects of larger-scale programs include analyses of the Chicago Child-Parent centers and some state pre-kindergarten programs (Reynolds 1998; Gormley and Gayer 2005; Wong et al. 2008). Cascio and Schanzenbach (2013) find small effects of programs in Georgia and Oklahoma for poor children, and no effects for richer children. Fitzpatrick (2008) finds small effects for Georgia's program, though some subgroups benefit.

[3] In other analyses of the HSIS data, Gelber and Isen (2013) show that Head Start participation increased parental involvement with children after the program ended, while Bitler et al. (2012) show that the reduced form impact of assignment to Head Start is larger for lower quantiles of the distribution of Peabody Picture and Vocabulary Test (PPVT) scores.

home visits, all of which have been cited as central to the success of model programs (Schweinhart 2007). The aim of this paper is to assess the contribution of these key inputs to cross-center differences in Head Start effects.

My analysis proceeds in two steps. First, to ask whether there is meaningful variation to be explained by program characteristics, I quantify heterogeneity in causal effects across Head Start centers. This investigation is complicated by non-compliance with random assignment in the HSIS experiment. Instrumental variables (IV) is the standard procedure for dealing with non-compliance, but IV has poor properties in samples the size of the applicant pools at individual Head Start centers (Nelson and Startz 1990). As a consequence, the conventional empirical Bayes "shrinkage" approach to quantifying effect heterogeneity, which requires unbaised estimates with known sampling variance, is invalid when applied to center-specific IV estimates.

I therefore develop a random coefficients sample selection model to measure cross-center variation in compliance probabilities and Head Start effects. The model combines a selection correction in the spirit of Heckman (1979) with a random coefficients structure to capture the distribution of treatment effects across experimental sites. This approach can be applied in more general settings to quantify heterogeneity in instrumental variables estimands across many groups. Estimates using this method reveal substantial heterogeneity in short-run Head Start effectiveness: the cross-center standard deviation of short-run cognitive effects is 0.18 test score standard deviations, larger than typical estimates of variation in teacher and school effectiveness (Deming 2014; Chetty et al. 2013a; Kane et al. 2008).

In a second step, I ask whether this variation can be explained by differences in program characteristics. My results show that some inputs play a role: Head Start centers offering full-day programs boost cognitive skills more than other centers, while centers offering frequent home visits are especially effective at raising non-cognitive skills. High/Scope Head Start centers are no more effective than other centers, however, and short-run effects are uncorrelated with teacher education, teacher certification, and class size. Together, observed inputs explain about 25 percent of the variation in Head Start cognitive effects. Finally, I investigate the role of counterfactual childcare choices, and show that cognitive gains are smaller for centers where a larger share of children would attend private preschool in the absence of the program. These findings suggests that replicating the effects of successful programs may be difficult, as the factors responsible for their success are largely unidentified.

The rest of the paper is organized as follows. The next section provides background on Head Start and describes the HSIS data. Section 3 summarizes the average impact of Head Start on summary indices of cognitive and non-cognitive skills. Section 4 outlines the random coefficients model used to investigate effect heterogeneity, and reports the results of this investigation. Section 5 analyzes the relationship between program characteristics and Head Start effectiveness. Section 6 concludes.

2 Data and Background

2.1 Head Start and the Head Start Impact Study

Head Start, the largest early-childhood program in the United States, enrolls roughly one million 3- and 4-year-old children at a cost of about $8 billion annually. Head Start awards grants to public, private non-profit, and for-profit organizations that provide childcare services to children below 130 percent of the Federal Poverty Line, though up to 35 percent of children attending a Head Start childcare center can be from households above this income threshold. Grantees are required to match at least 20 percent of federal Head Start funding. Head Start is based on a "whole child" model of school readiness that emphasizes non-cognitive social and emotional development in addition to cognitive skills. The grant-based nature of the program allows for a wide variety of childcare settings and practices, though all grantee agencies must meet a set of program-wide performance standards (US Department of Health and Human Services 2011; US Office of Head Start 2012).

The data used in my analysis come from the Head Start Impact Study (HSIS), a randomized evaluation of the Head Start program. The 1998 Head Start Reauthorization Act included a congressional mandate to determine the program's effects. As a result, the US Department of Health and Human Services (DHHS) conducted a nationally representative randomized controlled trial (DHHS 2010, 2012). The HSIS data includes information on 84 regional Head Start programs, 353 Head Start centers, and 4,442 children, each of whom applied to a sample Head Start center in Fall 2002. Sixty percent of applicants were randomly assigned the opportunity to attend Head Start ("treatment"), while the remaining applicants were denied this opportunity ("control"). Randomization took place at the Head Start center level; the HSIS data includes weights reflecting the probability of assignment for each child, which are used to adjust for these differences below.[4]

The HSIS sample includes two age groups, with 55 percent of students entering at age 3 and 45 percent entering at age 4. Three-year-old applicants could attend Head Start for up to two years before entering kindergarten, and three-year-olds assigned to the control group could re-apply to Head Start centers as four-year-olds the next year. Four-year-old applicants could attend for a maximum of one year. The data used here follow the treatment and control groups through 1st grade. DHHS (2010) provides a complete description of the HSIS experimental design and data collection procedures. The Online Appendix details the procedure used to construct my sample from the HSIS data.

2.2 Outcomes

The HSIS data include a large number of outcomes, collected for up to 4 years after random assignment. I organize these outcomes into summary indices of cognitive and non-cognitive skills. Table 1 lists the

[4]Some centers conducted multiple rounds of random assignment with differing admission probabilities, and the HSIS weights do not account for these differences. The discussion in DHHS (2010) suggests that any such differences are likely to be small, however.

outcomes included in each group. Cognitive outcomes include scores on the Peabody Picture and Vocabulary Test (PPVT) and several Woodcock Johnson III (WJIII) measures of cognitive ability. Non-cognitive outcomes, derived from parental surveys, include measures of social skills (making friends, hitting and fighting) and attention-span (concentration, restlessness). I exclude non-cognitive measures for which almost all respondents (90% or more) gave the same answer.[5]

Following Kling et al. (2007) and Deming (2009), I construct indices to summarize the impact of Head Start attendance across the outcomes listed in each column of Table 1. Specifically, I define the summary index

$$Y_i \equiv \frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{y_{i\ell}-\mu_\ell}{\sigma_\ell}\right),$$

where $y_{i\ell}$ is outcome ℓ for student i, and μ_ℓ and σ_ℓ are the control group mean and standard deviation of this outcome. I define outcomes so that positive signs mean better performance, and standardize them separately by year and age cohort.

2.3 Applicant Characteristics

Head Start applicants typically come from families with low socioeconomic status. This can be seen in the first column of Table 2, which presents mean demographic characteristics for the HSIS control group. The demographic variables come from a baseline survey of parents conducted in the Fall of 2002; parents of 3,577 HSIS applicants (81 percent) responded to this survey. Thirty-nine percent of mothers in the sample did not complete high school, and 17 percent of participants were born to teen mothers. The average household income in the sample is $1,576 per month.

To check experimental balance, column (2) of Table 2 shows coefficients from regressions of baseline characteristics on assignment to Head Start, weighting by the HSIS baseline child weights to adjust for differences in the probability of assignment across centers. The treatment/control differences in means are statistically insignificant for all baseline variables, and the joint p-value from a test of the hypothesis that assignment to Head Start is unrelated to all characteristics is 0.51, suggesting that random assignment was successful.[6]

The last two rows of Table 2 show the effects of assignment to Head Start on applicants' preschool choices. Applicants assigned to Head Start were 66 percentage points more likely to participate in the program than applicants from the control group in the first year after random assignment. Sixteen percent of students from the control group attended Head Start, most likely by applying to other nearby Head Start centers outside

[5]The HSIS data also includes measures of non-cognitive skills reported by teachers. I do not use these measures since they are unavailable for many children before kindergarten, and my analysis focuses on outcomes during preschool.

[6]Even with successful random assignment, non-random attrition has the potential to bias the experimental results. Appendix Table A1 shows attrition rates for the HSIS sample by year and outcome group, as well as treatment/control differences conditional on the controls included in Table 4. In preschool, outcomes are observed for 82 to 84 percent of children; the follow-up rate falls slightly in elementary school. Cognitive outcomes in preschool are observed slightly more frequently for children in the treatment group (3 percentage points) and this difference is marginally statistically significant. This modest differential attrition seems unlikely to drive the results reported below.

the experimental sample. Eighteen percent of children assigned to Head Start did not participate in the program. Together, these facts show that non-compliance with the experimental assignment is an important feature of the HSIS data, which motivates the instrumental variables approach taken below. The last row of Table 2 shows that a Head Start offer increases the probability of attending any center-based preschool program by 44 percentage points. This implies that two-thirds (0.442/0.663) of children induced to attend Head Start by the experimental offer would not have attended preschool otherwise, while the remaining one-third would have attended another preschool center if denied the opportunity to attend Head Start.

2.4 Center Characteristics

In addition to background information on applicants, the HSIS data includes detailed information on Head Start centers and their practices. I focus on inputs and practices that have been cited as central to the success of small-scale model programs. Schweinhart (2007) offers one view of the inputs that drove the success of the Perry Preschool Project:

> "The external validity or generalizability of the study findings extends to those programs that are reasonably similar to the High/Scope Perry Preschool Program. A reasonably similar program is a preschool education program run by teachers with bachelor's degrees and certification in education, each serving up to 8 children living in low-income families. The program runs 2 school years for children who are 3 and 4 years of age with daily classes of 2.5 hours or more, uses the High/Scope model or a similar participatory education approach, and has teachers visiting families at least every two weeks or scheduling regular parent events."

This account of the Perry program's effects emphasizes six key inputs: teacher education, teacher certification, class size, instruction time, the High/Scope curriculum, and home visiting. High/Scope is a participatory curriculum that emphasizes childrens' hands-on choices and experiences rather than adult-driven instruction (Epstein 2007). Schweinhart (2007) places particular weight on the High/Scope curriculum, arguing that results from the Perry Project and the followup High/Scope Preschool Curriculum Comparison Study "[suggest] that the curriculum had a lot to do with the findings."

No Head Start center replicates the Perry model, which used high levels of all six inputs and spent nearly three times as much as the average Head Start program on a per-pupil basis.[7] There is substantial variation in each of the six key Perry inputs within Head Start, however. This can be seen in Table 3, which summarizes characteristics of centers in the HSIS sample. One-third of Head Start centers use the High/Scope curriculum. Thirty-five percent of Head Start teachers have bachelor's degrees, and 12 percent hold teaching licenses, but the fractions with these credentials range from zero to 100 percent across centers. The average Head Start center has 6.8 children for every staff member; the cross-center standard deviation

[7] Heckman et al. (2010a) report that the Perry program cost about $18,000 per child in 2006 dollars. Per-child expenditure in Head Start was $7,600 in 2011, which is $6,800 deflated to 2006 dollars using the Consumer Price Index series available at http://www.bls.gov (DHHS 2011).

of class size is 1.7 children. Sixty-three percent of Head Start centers provide full-day service, and 20 percent offer more than three home visits per year. In Section 5, I explore whether this variation in inputs can explain differences in effectiveness across Head Start centers.

3 LATE Framework and Pooled Estimates

3.1 LATE Framework

I next describe the local average treatment effect (LATE) framework used to interpret IV estimates of Head Start effects (Imbens and Angrist 1994). Consider a set of students, each applying to Head Start at one of J students. Let $C_i \in \{1, ..., J\}$ indicate the center where student i applies. The Bernoulli variable Z_i indicates random assignment to Head Start, and D_i indicates Head Start attendance. Let $D_i(1)$ and $D_i(0)$ indicate potential Head Start attendance as a function of Z_i, so that $D_i = D_i(Z_i)$. Finally, there is an outcome variable, Y_i, with potential values $Y_i(1)$ and $Y_i(0)$ indexed against D_i. The observed outcome is $Y_i = Y_i(D_i)$.

The LATE framework is built on the following assumptions:

1. Independence/Exclusion: $(Y_i(1), Y_i(0), D_i(1), D_i(0))$ is independent of Z_i conditional on C_i
2. First Stage: $Pr\left[D_i = 1|Z_i = 1, C_i = j\right] > Pr\left[D_i = 1|Z_i = 0, C_i = j\right] \ \forall j$
3. Monotonicity: $Pr\left[D_i(1) \geq D_i(0)|C_i = j\right] = 1 \ \forall j$

Assumption (1) follows if $Z_i$is randomly assigned within centers and only affects outcomes through Head Start attendance. Assumption (2) requires that assignment to Head Start induces some children to attend Head Start at every center. Assumption (3) requires that assignment to Head Start does not discourage any child from participating in the program.

Under these assumptions, the population can be partitioned into three groups defined by response to Head Start assignment. *Compliers* participate in Head Start if assigned to the program and not otherwise $(D_i(1) > D_i(0))$. *Never takers* decline to participate regardless of assignment $(D_i(1) = D_i(0) = 0)$. *Always takers* participate even if assigned to the control group $(D_i(1) = D_i(0) = 1)$. Imbens and Angrist (1994) show that instrumental variables estimates recover local average treatment effects (LATE), average causal effects of Head Start attendance for compliers. We have

$$\frac{E\left[Y_i|Z_i = 1, C_i = j\right] - E\left[Y_i|Z_i = 0, C_i = j\right]}{E\left[D_i|Z_i = 1, C_i = j\right] - E\left[D_i|Z_i = 0, C_i = j\right]} = E\left[Y_i(1) - Y_i(0)|D_i(1) > D_i(0), S_i = j\right] \tag{1}$$

$$\equiv \beta_j.$$

β_j is the LATE at center j. The Wald (1940) instrumental variables estimator replaces population moments with sample moments on the left-hand side of equation (1).

3.2 Pooled Estimates

Before investigating heterogeneity in causal effects, I summarize the average impact of Head Start using pooled equations of the form

$$Y_i = \alpha + \beta D_i + X_i'\lambda + \epsilon_i, \tag{2}$$

where Y_i is a summary index of outcomes for student i,D_i is a dummy for Head Start attendance, and X_i is a vector of the baseline controls from Table 2, included to increase precision. The attendance dummy is instrumented with an indicator for assignment to Head Start, Z_i, with first stage equation

$$D_i = \kappa + \pi Z_i + X_i'\delta + \eta_i. \tag{3}$$

I estimate these equations by weighted two-stage least squares using the HSIS baseline child weights to account for differences in the probability of assignment across centers. The coefficient β can be interpreted as a a weighted average of center-specific LATEs (Angrist and Imbens 1995).[8]

Estimates of equations (2) and (3) reveal that Head Start attendance boosts outcomes during preschool, but these effects fade out quickly once children leave the program. Table 4 reports estimates of effects for cognitive and non-cognitive skills, separately by grade and assignment cohort. Column (1) shows that in the first year after random assignment, applicants assigned to treatment were 68 percentage points more likely to attend Head Start than applicants in the control group. The corresponding second-stage estimates for cognitive skills, reported in column (2), shows that Head Start attendance increased cognitive skills by 0.17 standard deviations for three-year-olds and 0.09 standard deviations for four-year-olds. These estimates are statistically significant at the 5-percent level, and (when appropriately adjusted for the first stage) they are consistent with intent-to-treat effects for individual outcomes reported by DHHS (2010) and Bitler et al. (2012). Estimates for non-cognitive skills, reported in column (4), show smaller effects. The point estimate for three-year-olds is positive and marginally statistically significant, but the estimate for four-year-olds is negative and statistically insignificant.

In Spring 2004, members of the three-year-old cohort were still enrolled in Head Start. The cognitive point estimate for this time period is comparable to the Spring 2003 estimate (0.16 standard deviations), but is less precise (s.e. = 0.078). The decline in precision between 2003 and 2004 is driven by a decline in compliance for the three-year-old cohort: many children in the control group re-applied to Head Start and were admitted at age 4, reducing the first stage from 0.68 to 0.35.[9] Similarly, the non-cognitive estimate for three-year-olds in Spring 2004 is positive, but imprecise.

The remaining rows of Table 4 show that the effects of Head Start attendance dissipate once children exit

[8] Angrist and Imbens (1995) show that two-stage least squares estimation of a system using all center-by-treatment interactions as instruments produces a weighted average of center-specific LATEs, with weights proportional to the variance of the first stage fitted values. Estimates from this saturated model were similar to weighted least squares estimates of equations (2) and (3).

[9] Head Start participation in Spring 2004 is measured from the parental survey since an administrative measure of participation is only available in Spring 2003. See the Online Appendix.

the program. Cognitive estimates for both cohorts are statistically insignificant in kindergarten (Spring 2005 for three-year-olds, Spring 2004 for four-year-olds). The estimates for four-year-olds are precise; effect sizes smaller than 0.1 standard deviations can be rejected at the 5-percent confidence level for both cognitive and non-cognitive skills. Estimates for both three- and four-year-olds are also small and statistically insignificant in 1st grade. Non-cognitive estimates for three-year-olds are positive in Spring 2004, Spring 2005, and Spring 2006, but these estimates are not significantly distinguishable from zero. Together, these results show little evidence of cognitive or non-cognitive effects once children leave preschool.

4 Variation in Head Start Effects

4.1 Variation in Instrumental Variables Estimates

I next turn to the primary contribution of this paper: Quantifying and explaining variation in causal effects across Head Start centers. As a first look at cross-center heterogeneity, Figure 2 plots center-specific reduced form coefficients against first stages. These coefficients come from regressions of cognitive skills and Head Start attendance on the Head Start offer indicator, respectively. In the absence of treatment effect heterogeneity, reduced forms should be proportional to first stages with the same constant of proportionality for every center, so a single line through the origin should fit all points in Figure 2 up to sampling error. The red lineshows a weighted least squares regression through the origin, with weights proportional to sample size times the variance of the Head Start offer. The χ^2 statistic from a test that all points line on this line is equal to the overidentification test statistic from a two-stage least squares model using all center-by-offer interactions as instruments for Head Start attendance. The χ^2 statistic is equal to 424.9 and has 320 degrees of freedom, so the null hypothesis of no cross-center effect heterogeneity is rejected ($p < 0.01$).

The evidence in Figure 3 suggests that effects vary across Head Start centers. The magitude of this variation is also of interest. Parametric empirical Bayes (EB) methods are the conventional approach to quantifying cross-site variation in treatment effects (Morris 1983). The EB approach involves specifying a prior distribution for the cross-site distribution of parameters, and then estimating the hyperparameters of the prior. In standard cases where site-specific estimates are unbiased and have a known sampling variance, the EB estimator takes an especially simple form: The variance of treatment effects can be consistently estimated by subtracting the average squared standard error from the sample variance of site-specific estimates (Jacob and Lefgren 2008). An efficient "shrinkage" estimator of the effect at a particular site can then be constructed as a weighted average of the estimate for that site and the overall average effect.

This simple approach is inappropriate for the HSIS data. To account for differences in compliance with random assignment across centers, it is necessary to study instrumental variables estimates rather than intent-to-treat effects of assignment to Head Start. Instrumental variables estimates have no finite moments and are not centered at the true parameter in finite samples (Nelson and Starz 1990). In addition, conventional asymptotic standard errors are likely to provide a poor approximation to their behavior in small samples

(Mariano 1977). Center-specific samples in the HSIS are often small, so the finite-sample behavior of IV is relevant for center-specific IV estimates. This can be seen in Figure 1, which shows a histogram of the distribution of sample sizes across HSIS centers. More than half of centers have fewer than 10 applicants, and few have more than 25.

Table 5 illustrates the poor finite-sample behavior of center-specific IV estimates for cognitive skills in Spring 2003. The IV estimate for center j, $\hat{\beta}_j$, is the sample analogue of equation (1). The sample standard deviation of these estimates is large (1.39 test score standard deviations), and estimates for some centers are implausible (as large as 15.2 standard deviations). The wide dispersion in center-specific estimates is evident in Figure 3, which shows a histogram of $\hat{\beta}_j$, excluding estimates in excess of 2 in absolute value to keep the scale reasonable. Moreover, the asymptotic standard errors associated with these estimates yield nonsensical results. The average standard error is 6.2 standard deviations. As a result of extremely large standard errors for some centers, a simple estimate of the variance of β_j formed by subtracting the average squared standard error from the sample variance of $\hat{\beta}_j$ yields a large negative number.[10]These results make it clear that the $\hat{\beta}_j$ and their asymptotic standard errors are not informative about the extent of effect heterogeneity across centers. I next describe a framework that consistently quantifies variation in Head Start effects despite small within-center sample sizes.

4.2 Random Coefficients Framework

My approach to quantifying effect variation uses a sample selection model to describe potential oucomes and Head Start participation conditional on Z_i and center-specific parameters. I treat the parameters at each center as draws from a prior distribution of random coefficients, and derive an integrated likelihood function for the sample that depends only on the hyperparameters of this distribution. I then estimate the hyperparameters by maximum likelihood. This approach circumvents the need to compute $\hat{\beta}^j$ for every Head Start center.

Potential outcomes at center j can be written

$$Y_{ij}(d) = \alpha_{dj} + \epsilon_{idj},\ d \in \{0, 1\},$$

where the subscript j now refers to a student's center of random assignment and $E[\epsilon_{idj}] = 0$. The Head Start participation decision is described by

$$D_{ij} = 1\{\lambda_j + \pi_j Z_{ij} > \eta_{ij}\}.$$

The vector of parameters at center j is therefore

$$\theta_j \equiv (\alpha_{1j}, \alpha_{0j}, \lambda_j, \log \pi_j)'.$$

[10]I also used the iterative procedure for producing restricted maximum likelihood estimates of effect variation suggested by Morris (1983). This procedure places less weight on observations with large estimated sampling variances. The iterative procedure failed to converge and produced a variance estimate arbitrarily close to zero.

The average treatment effect (ATE) of Head Start attendance at center j is $\alpha_{1j} - \alpha_{0j}$. Note that the parameter vector is defined in terms of $\log \pi_j$, which guarantees that a Head Start offer weakly increases the probability of Head Start participation for any value of θ_j. This preserves the monotonicity assumption of the LATE model.

I assume the following parametric structure for the within-center distribution of potential outcomes:

$$(\epsilon_{i1j}, \epsilon_{i0j}, \eta_{ij})' \sim N(0, \Sigma). \tag{4}$$

Conditional on the center-specific parameters θ_j, assumption (4) yields a two-sided version of the Heckman (1979) sample selection model. The likelihood of the observed outcomes for student i is given by

$$\mathcal{L}_{ij}(Y_{ij}, D_{ij}|Z_{ij}; \theta_j) = \left[\Phi\left(\frac{\sigma_1(\lambda_j + \pi_j Z_{ij}) - \rho_1(Y_{ij} - \alpha_{1j})}{\sigma_1\sqrt{1-\rho_1^2}}\right)\frac{1}{\sigma_1}\phi\left(\frac{Y_{ij} - \alpha_{1j}}{\sigma_1}\right)\right]^{D_{ij}} \tag{5}$$

$$\times\left[\left(1 - \Phi\left(\frac{\sigma_0(\lambda_j + \pi_j Z_{ij}) - \rho_0(Y_{ij} - \alpha_{1j})}{\sigma_0\sqrt{1-\rho_0^2}}\right)\right)\frac{1}{\sigma_0}\phi\left(\frac{Y_{ij} - \alpha_{0j}}{\sigma_0}\right)\right]^{1-D_{ij}},$$

where σ_d is the standard deviation of ϵ_{ijd} and ρ_d is its correlation with η_{ij}.

To check whether assumption (4) is reasonable in the HSIS data, Table 6 compares IV estimates of LATE to maximum likelihood estimates from the normal selection model for Spring 2003, restricting parameters to be the same across centers ($\theta_j = \theta_0\ \forall j$). Column (5) shows estimates of the ATE, while column (4) shows the LATE implied by the selection model.[11]

The full set of maximum likelihood estimates is listed in Table A2. The results in Table 6 show that the maximum likelihood results closely match the IV estimates. The selection model exactly reproduces the first-stage effect of a Head Start offer, and the maximum likelihood estimates of LATE in column (4) are almost indistiguishable from the IV estimates in column (2). Columns (4) and (5) also show that estimates of ATE and LATE are very similar. As can be seen in Table A2, this is due to the fact that estimates of ρ_1 and ρ_0 are both around 0.12. The estimated correlations are significantly different from zero, which suggests the presence of endogenous selection into Head Start; however, with $\rho_1\sigma_1 \approx \rho_0\sigma_0$, treatment effects are unrelated to the propensity to take up the treatment, so LATE and ATE are approximately the same. This suggests the absence of what Heckman et al. (2006) term "essential heterogeneity," treatment effect heterogeneity that is systematically related to the propensity to participate in Head Start.

Next, I assume that the cross-center distribution of parameters follows a prior distribution F:

$$\theta_j|Z_j \sim F(\theta; \Omega).$$

[11] The LATE is given by

$$LATE = \alpha_1 - \alpha_0 + E[\epsilon_{i1j} - \epsilon_{i0j}|\lambda < \eta_{ij} < \lambda + \pi]$$

$$= \alpha_1 - \alpha_0 + (\rho_1\sigma_1 - \rho_0\sigma_0)\cdot\left(\frac{\phi(\lambda) - \phi(\lambda+\pi)}{\Phi(\lambda+\pi) - \Phi(\lambda)}\right).$$

The hyperparameter Ω captures heterogeneity in outcome distributions and experimental compliance between Head Start centers. To estimate Ω, I integrate the site-specific parameters out of the likelihood function. The integrated likelihood for center j is

$$\mathcal{L}_j^I\left(Y_j, D_j | Z_j; \Omega\right) = \int \prod_i \mathcal{L}_{ij}\left(Y_{ij}, D_{ij} | Z_{ij}; \theta\right) dF\left(\theta; \Omega\right). \tag{6}$$

An empirical Bayes (EB) estimate of Ω maximizes the sum of logarithms of integrated likelihoods across Head Start centers.

I estimate two versions of the model with different specifications for the prior distribution F. In the first, $\theta_j | Z_j$ is assumed to follow a multivariate normal distribution with mean θ_0 and covariance matrix V_0. In the second, Head Start centers are assumed to belong to a finite set of K possible types, so that $\theta_j \in \left\{\theta^1, \theta^2, ..., \theta^K\right\}$; type probabilities are given by $Pr\left[\theta_j = \theta^k | Z_j\right] = P^k$. The integral in equation (6) does not have a closed form in the normal case, so I approximate it by simulation, using 1,000 draws of θ_j for each center. [12]

4.3 Random Coefficients Estimates

Table 7 reports key parameter estimates from the normal random coefficients model for Spring 2003, pooling the three- and four-year-old cohorts.[13] Additional parameter estimates are reported in Appendix Table A3. I focus on Spring 2003 because effects for this period are largest and precisely estimated; in addition, the evidence in Chetty et al. (2011) suggests that immediate impacts of early-childhood programs may predict long-run effects better than impacts in later time periods. Results for Spring 2005 are reported in Appendix Tables A3 and A4.

The estimated random coefficient distributions reveal significant heterogeneity in parameters across Head Start centers. Consistent with the first stage estimates in Table 6, the mean compliance probability is 0.75. Compliance rates vary substantially across sites: The cross-site standard deviation of the compliance probability is 0.21 . This implies that about 20 percent of centers have compliance probabilities below 0.5.

Table 7 also shows estimates of the cross-center distribution of causal effects. The estimate of the average effect for cognitive skills is 0.12 standard deviations, while the mean non-cognitive effect is 0.03. Encouragingly, these mean impacts are similar to the pooled estimates in Table 6. The cross-center standard deviation of Head Start effects, given by $\sqrt{Var(\alpha_{1j} - \alpha_{0j})}$, is estimated to be 0.18 standard deviations for cognitive skills. These implies substantial treatment effect variation across Head Start centers. For

[12] The simulated likelihood for center j is given by $\tilde{\mathcal{L}}_j^I\left(Y_j, D_j | Z_j; \Omega\right) = \frac{1}{R}\sum_{r=1}^{R}\prod_i \mathcal{L}_{ij}\left(Y_{ij}, D_{ij} | Z_{ij}; \theta_j^r(\Omega)\right)$, where R is the number of draws and $\theta_j^r(\Omega)$ is the r-th draw for center j. The simulated maximum likelihood estimator has the same asymptotic distribution as the conventional maximum likelihood estimator as long as R rises faster than $\sqrt{J}$ (Train 2003).

[13] Within a center, three- and four-year-old applicants sometimes faced different probabilities of assignment to Head Start. I reweight likelihood contributions to account for these differences. Specifically, the likelihood contribution of child i is $\mathcal{L}_{ij}^{w_i}$, where $\mathcal{L}_{ij}$ is the expression for the likelihood given in equation (5) and w_i is a weight proportional to child i's base HSIS weight, normalized to sum to the total sample size.

comparison, estimates of the standard deviations of school and teacher effectiveness are typically around 0.1 test score standard deviations (Chetty et al. 2013a; Deming 2014; Kane et al. 2008). My estimates suggest that variation in short-run Head Start effectiveness is nearly twice as large as variation in value-added across teachers or schools. The standard deviation of effects for non-cognitive skills is smaller (0.066σ). Figure 2 summarizes the estimated prior distributions for the normal model, comparing them to histograms of center-specific first stage and IV estimates. The estimated prior distributions show much less dispersion than the distributions of center-specific estimates, particularly for non-cognitive skills; nonetheless, the priors display substantively important heterogeneity.

As an alternative to the normal model, Appendix Table A5 reports estimates from models assuming that Head Start centers belong to a finite set of discrete types. The finite-type estimates also suggest significant heterogeneity in cognitive effects across Head Start centers. Estimates from a three-type model, shown in columns (1) through (3), reveal two types of centers with high compliance rates and cognitive effects around 0.13 standard deviations. Type 1 centers have relatively low scores in both the treated and non-treated states, while type 2 centers have higher scores. The third type of center is less common (13 percent of centers), and has a negative treatment effect (-0.21σ) and a much lower compliance rate. The implied cross-center standard deviation of effects is 0.12 standard deviations, somewhat smaller than the estimate from the normal model.

Columns (4) through (8) report estimates of a model that allows five types of Head Start center. The majority type has a compliance rate of 0.75 and a treatment effect of 0.11, similar to the three-type case. The remaining four types are very heterogeneous. Type 2 has high mean scores and a treatment effect ear zero, while type 3 has lower scores in the control state and a very large effect (0.55 standard deviations). Type 4 has a significant negative effect, but has a small population share (3 percent); type 5 is also rare (9 percent) and a compliance rate very close to zero. The five-type model produces a cross-center standard deviation of treatment effects of 0.22 standard deviations, slightly larger than the corresponding estimate from the normal model. Estimates of models with more than five types produce vary small population shares for the additional types, suggesting that the five-type model does a reasonable job of capturing cross-center heterogeneity. Together, the normal and finite-type random coefficient estimates both suggest significant variation in effects across Head Start centers.

To provide further context for the random coefficient estimates, I next compute the implied earnings effect of an improvement in Head Start quality, using the relationships between test score effects and lifetime earnings reported by Chetty et al. (2013b). Chetty et al. (2013b) show that a one-standard-deviation increase in teacher value-added in a single grade translates into a 1.3 percent increase in lifetime earnings. If the mapping between the short-run effect of Head Start on test scores and its effect on earnings is the same as this mapping for teachers, my results imply that a Head Start center at the 84th percentile of program quality (one standard deviation above average) will boost lifetime earnings by 1.8 percent relative to the average Head Start center. Assuming that children in the HSIS data will earn roughly the same amount as

their parents relative to the national median (a conservative assumption since earnings revert to the mean), and using the same the assumptions on lifetime earnings trajectories described by Chetty et al. (2013b), this translates into an earnings effect of about $3,400 per child in 2010 dollars.[14] This calculation shows that the magnitude of cross-center variation in Head Start effectiveness is large enough to matter for later outcomes, and is also large relative to the per-child cost of the program (roughly $7,600; DHHS 2011).

5 Explaining Head Start Effects

5.1 Center Characteristics

The estimates reported above show that some Head Start programs are substantially more effective than others. In the remainder of the paper, I ask whether this variation in effectiveness can be explained by observed inputs. The analysis of inputs focuses on the six key variables cited by Schweinhart (2007) as responsible for the success of the Perry Preschool Project: The High/Scope curriculum, teacher education and certification, class size, instructional time, and home visting.

I investigate the relationship between inputs and causal effects using two approaches. First, I estimate interacted two-stage least squares models, with second- and first-stage equations of the form

$$Y_{ij} = \alpha + P_j'\phi + \beta D_{ij} + D_{ij} \cdot P_j'\psi + \epsilon_{ij}, \tag{7}$$

$$D_{ij} = \kappa + P_j'\nu + \pi Z_{ij} + Z_{ij} \cdot P_j'\rho + \eta_{ij}, \tag{8}$$

where P_j is a vector of inputs and practices at center j. The first stage equations for the interactions of D_{ij} and P_j are analogous to equation (5). This approach compares IV estimates for centers with different values of P_j. The vector ψ captures the relationship between the effect of Head Start attendance and observed inputs.

Second, I extend the normal random coefficients model to incorporate dependence between inputs and causal effects in the prior distribution. Specifically, I write center-specific parameters as

$$\theta_j = \theta_0 + \Gamma P_j + \xi_j,$$

where $\xi_j \sim N(0, V_0)$. The effect of Head Start attendance at center j is then

[14] Chetty et al. (2013b) report that the standard deviation of teacher quality is 0.13 test score standard deviations. They argue that a one-standard-deviation move upwards in this teacher quality distribution for one year raises students' earnings by 1.3 percent. The implied earnings gain per standard deviation of test scores is therefore (1.3/0.13) = 10 percent. I estimate that the standard deviation of Head Start quality is 0.18 test score standard deviations, so a one standard deviation increase in Head Start quality boosts earnings by 0.18·10 = 1.8 percent. Chetty et al. (2013b) estimate that the mean present value of lifetime earnings is roughly $522,000 at age 12 in 2010 dollars, which is $434,000 discounted back to age 5 at a 3-percent rate. The average HSIS family earned $18,912 per year, or 45 percent of the US median in 2002 (see http://www.census.gov/prod/2003pubs/p60-221.pdf). The average present discounted value of earnings at age 5 for children in the HSIS sample can therefore be conservatively estimated as $0.44 \cdot \$434,000 = \$190,960$. The earnings impact of a 1 standard deviation increase in Head Start quality can then be approximated as $190,960·0.018= $3,437.28.

$$\beta_j = (\Gamma^1 - \Gamma^2)P_j + \left(\xi_j^1 - \xi_j^2\right),$$

where Γ^k is the k-th row of Γ and similarly for ξ_j^k. This approach relies in part on the parametric assumptions described in Section 4, so it is likely to be less robust than two-stage least squares. The advantage of the random coefficients approach is that it generates an estimate of V_0, the residual variation in center-specific parameters remaining after accounting for observed inputs. It can therefore be used to measure the share of effect heterogeneity explained by P_j.

Table 8 reports estimates using these two approaches. I estimate two sets of interaction models: bivariate models that include inputs in P_j one at a time, and multivariate models that include all six inputs in P_j. The estimates in columns (1) through (3) reveal that centers providing full-day service have larger cognitive effects than other centers. On average, effects of full-day centers are 0.13 standard deviations larger than effects of centers that do not offer this service. Estimates for the multivariate interaction and maximum likelihood models are similar. This implies that the relative effectiveness of full-day centers is not explained by other inputs.

The remaining rows of Table 8 show that the other five inputs are uncorrelated with cognitive effects in Head Start. High/Scope centers do not boost scores more than non-High/Scope centers; the interaction terms associated with High/Scope are very close to zero in all models. Moreover, this difference is reasonably precisely estimated: I can reject the hypothesis that High/Scope centers are 0.12 standard deviations more effective than other centers at the 5-percent confidence level. This result weighs against the view that the High/Scope curriculum alone generated much of the success of the Perry Preschool Project (Schweinhart 2007).

Estimates of differences in cognitive effects associated with teacher education, teacher licensing, class size, and home visiting are also statistically insignificant. In the bivariate model, the interaction coefficient for the share of teachers with bachelors degrees is -0.012, with a standard error of 0.076. The upper bound of the 95-percent confidence interval for the effect of a one-standard-deviation increase in teacher education (a 40 percentage point change) is therefore $(-0.012 + 0.076 \cdot 1.96) \cdot 0.4 = 0.055$ standard deviations, a relatively small effect. Similarly, the interaction coefficients associated with teacher certification are insignificant in all model. These results are consistent with previous evidence showing that teacher effectiveness is difficult to predict with observed characteristics (Kane et al. 2008).

The results for student/staff ratios are more surprising. Estimates from both experimental and quasi-experimental settings suggest that reduced class size boosts test scores (Chetty et al., 2011; Krueger, 1999; Angrist and Lavy 1995). In contrast, the results in Table 8 suggest that Head Start centers with larger classes are not less effective. The estimated relationship between frequency of home visiting and cognitive effects is statistically insignificant, but imprecise enough that relatively large effects cannot be ruled out.

Columns (4) through (6) of Table 8 reveal a broadly similar pattern for non-cognitive skills, with a few important differences. The High/Scope curriculum, teacher characteristics, and class size are not correlated with non-cognitive effectiveness. Unlike the cognitive estimates, however, the non-cognitive estimates for

full-day service are small and insignificant. In addition, though it is not correlated with cognitive effects, frequency of home visiting seems to be related to non-cognitive effectiveness in Head Start. The two-stage least squares estimates show that centers offering more than three home visits per year boost non-cognitive skills by 0.13 standard deviations more than centers providing three or less visits. The corresponding maximum likelihood estimate is 0.09 standard deviations. While it is not possible to unpack the mechanisms responsible for differences in cognitive and non-cognitive effects, one possibility is that cognitive effectiveness may be determined more by classroom-level factors (e.g. instructional time), while non-cognitive effectiveness may be influenced more by factors that alter the quality of the home environment (e.g. home visiting).

The final row of Table 8 reports estimates of $\sqrt{Var\left(\xi_j^1 - \xi_j^2\right)}$, the residual standard deviation of Head Start effects after accounting for observed inputs. Residual standard deviations are 0.154 for cognitive skills and 0.056 for non-cognitive skills. These estimates are only slightly smaller than the corresponding standard deviations in Table 7, which implies that inputs explain a small share of the variation Head Start effects. More specifically, in an R^2 sense, inputs explain 28 percent of the variation in cognitive effects, and 29 percent of the variation in non-cognitive effects.[15] Most of the heterogeneity in impacts across Head Start centers is therefore unexplained by the six inputs emphasized by Schweinhart (2007).

It is important to note that inputs and practices are not randomly assigned to Head Start centers, so the estimates in Table 8 may not reflect causal impacts of changing inputs in isolation. If the inputs analyzed here are correlated with unobserved center-level factors that influence effectiveness, these estimates will provide a misleading picture of the effects of changing inputs. Moreover, there may be important complementarities between inputs, which are not accounted for in the linear models for treatment effects implied by equations (7) and (8). Nonetheless, this analysis provides little evidence that adoption of the High/Scope curriculum or teacher credentialing requirements would improve program effectiveness in Head Start. This finding is relevant to recent policy changes that mandate increased education levels for Head Start teachers (DHHS 2008). My results show that full-day service and home visiting are most predictive of short-run Head Start effectiveness.

5.2 Counterfactual Preschool Choices

In this section, I explore an alternative explanation for heterogeneity in effects across Head Start centers: variation in preschool choices for children that do not attend Head Start. Children in the HSIS sample can participate in three types of childcare: Head Start, private center-based preschool, or home care (no preschool). As shown in Table 2, the effect of a Head Start offer on the probability of Head Start attendance is larger than its effect on preschool attendance; this implies that some applicants would attend other preschools in the absence of Head Start. If private preschool affects cognitive skills relative to no preschool, differences in private preschool participation rates may drive cross-center variation in Head Start effects even if Head

[15] The proportions of variation in cognitive and non-cognitive effects explained by inputs are $1 - \left(\frac{0.154}{0.181}\right)^2$ and $1 - \left(\frac{0.056}{0.066}\right)^2$.

Start programs have uniform quality.

To investigate this issue, I estimate the share of students drawn into Head Start from private preschool at center j using the regression

$$PR_{ij} = \tau_j^{PR} + \rho_j^{PR} Z_{ij} + u_{ij}^{PR}. \tag{9}$$

where PR_{ij} is an indicator for private preschool attendance. The coefficient ρ_j^{PR} measures the total reduction in private preschool attendance caused by a Head Start offer. Similarly, the share of students drawn from no preschool can be estimated using the regression

$$N_{ij} = \tau_j^{N} + \rho_j^{N} Z_{ij} + u_{ij}^{N}, \tag{10}$$

where N_{ij} is an indicator for attending no preschool. Under the assumption that a Head Start offer does not affect the choice of private vs. no preschool,[16] the share of Head Start compliers drawn from no preschool is given by

$$S_j^N = \frac{\left(-\rho_j^N\right)}{\left(-\rho_j^N\right) + \left(-\rho_j^{PR}\right)}.$$

I estimate equations (9) and (10) by weighted least squares using the HSIS child weights, setting positive coefficients to zero to keep S_j^N beween zero and one. Figure 5 shows a histogram of S_j^N. This figure reveals that the share of compliers who would attend no preschool in the absence of Head Start varies across centers. At about 10 percent of centers, all compliers attend private preschool if denied the opportunity to attend Head Start. About twenty percent of centers appear to draw children only from home care. The remaining 70 percent draw children from a mix of private preschool and no preschool.

To explore the role of private preschool, I estimate interacted two-stage least squares models that allow the effect of Head Start to differ between students applying to centers above and below the sample median of S_j^N. Table 9 shows that the short run cognitive effect of Head Start is larger at centers that draw more students from no preschool. Above-median centers boost scores by 0.18 standard deviations. The corresponding effect for below-median centers is 0.06 standard deviations, and equality of these effects is rejected at conventional levels ($p = 0.02$). Figure 6 plots corresponding estimates that split the sample by quartiles of S_j^N. The estimated effects are not monotonic, but the cognitive estimates for the two higher quartiles of S_j^N are larger than the corresponding effects for the lower quartiles. The share of compliers draw from no preschool seems to be unrelated to non-cognitive effects, however. While the relationships between complier shares and Head Start effects are not precisely estimated, the results suggest that some of the variation in effects across Head Start centers may be driven by differences in private preschool participation rates rather than characteristics of the centers themselves.

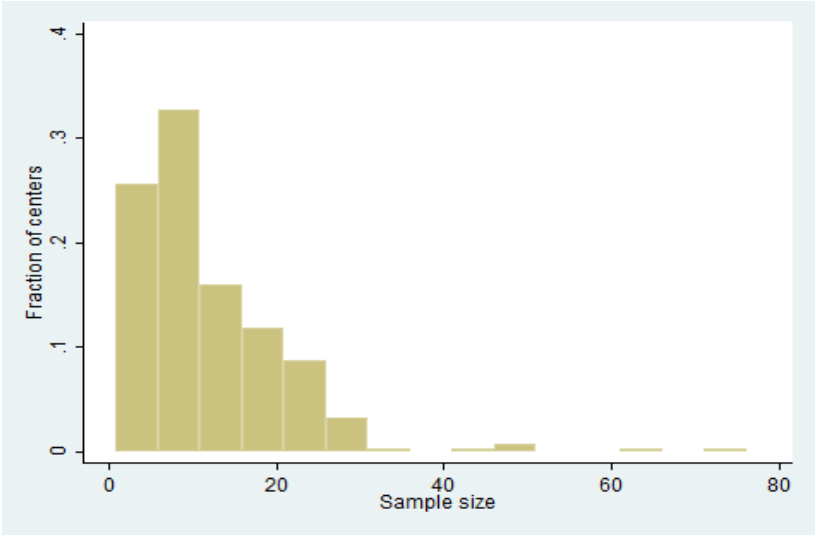
[16] This assumption can be motivated by a revealed preference argument: The availability of private preschool is unaffected by a Head Start offer, so preferences for private vs. no preschool should not be affected by the offer. A shift between private and no preschool in response to a Head Start offer would violate the exclusion restriction required for the offer to be a valid instrument for Head Start attendance.

6 Conclusion

Studies of small-scale model early-childhood programs show that early intervention can dramatically boost outcomes in the short- and long-run. Randomized evidence from the Head Start Impact Study (HSIS) suggests that the Head Start program produces smaller short-run gains. This paper uses data from the HSIS to quantify impact variation across Head Start centers and ask whether differences in key inputs used by model programs can explain this variation. A random coefficients instrumental variables analysis reveals substantial variation in effectiveness across Head Start centers, particularly with respect to cognitive skills. Centers with full-day service and with frequent home visiting are more effective than other centers, but other inputs typically cited as important to the success of small-scale programs, including the High/Scope curriculum, teacher education and certification, and class size, do not predict program effectiveness in Head Start. Together, these inputs explain less than 30 percent the variation in short-run cognitive effects across Head Start centers. An investigation of the role of counterfactuals suggests that variation in Head Start effects may be partially driven by differences in private preschool participation among children denied the opportunity to attend Head Start.

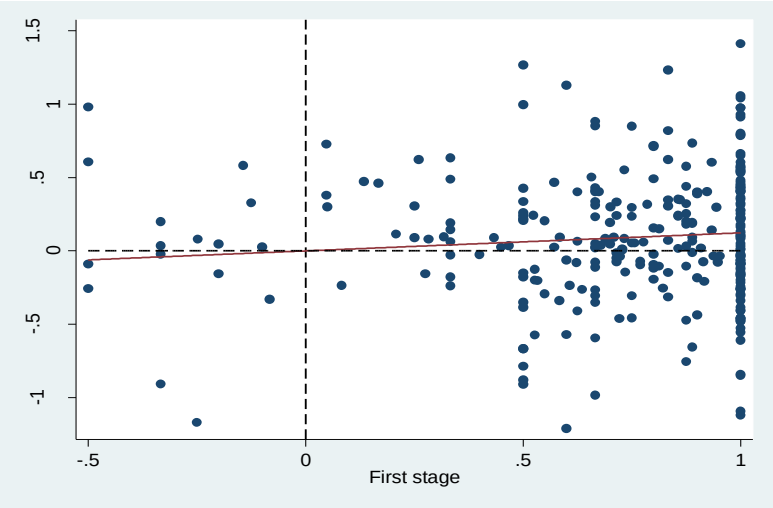
This analysis raises the further question of how the success of small-scale model programs can be replicated. This question is of immediate policy importance given recent calls to expand public preschool access (Obama 2013). The results reported here are similar to findings from the literature on teacher quality, which typically show weak relationships between observed teacher characteristics and value-added (Kane et al. 2008). These results suggest that replicating the effects of small-scale programs at a lower cost may be difficult, as the success of these programs may be due to unobserved inputs that are not easy to reproduce more cheaply and on a larger scale.

Figure 1: Histogram of Sample Sizes Across Head Start Centers



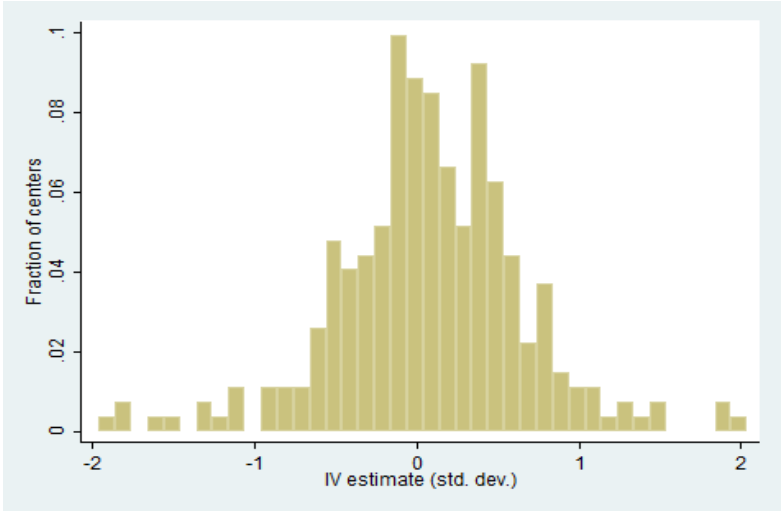
Notes: This figure shows the histogram of center-specific sample sizes in the HSIS experiment. The data are grouped into bins of 5 children (0 to 5, 6 to 10, etc.).

Figure 2: Center-specific Reduced Forms and First Stages



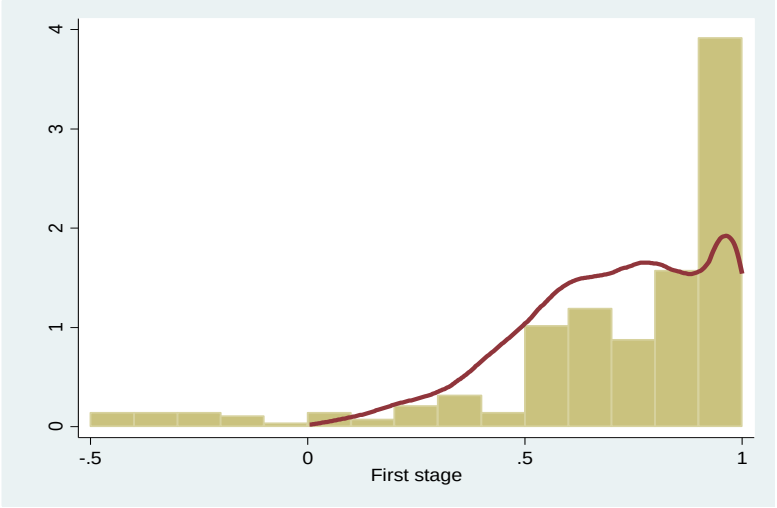
Notes: This figure plots center-specific reduced form differences in cognitive skills in Spring 2003 against first stage differences in Head Start attendance rates. The red line comes from a weighted least squares regression through the origin, with weights proportional to $NP(Z)[1 - P(Z)]$, where N is sample size and $P(Z)$ is the fraction of applicants offered Head Start. The slope is 0.12 (SE = 0.03). The $\chi^2(320)$ statistic from a test that all points line on the line is 424.9 ($p = 0.00$).

Figure 3: Histogram of Center-specific IV Estimates

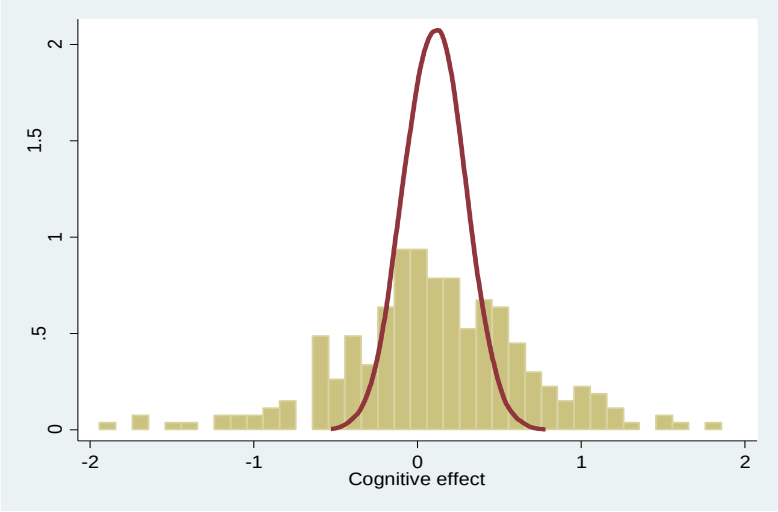


Notes: This figure plots the histogram of center-specific IV estimates for cognitive skills in Spring 2003. Estimates greater than 2 in absolute value are excluded.

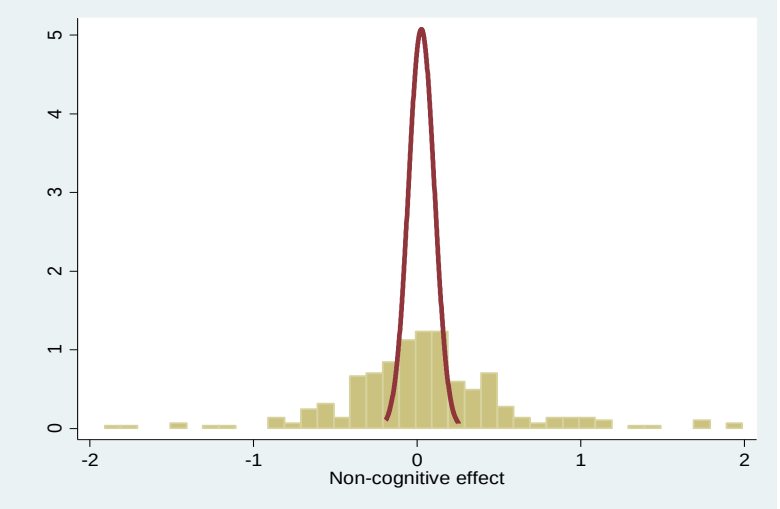
Figure 4: Empirical Bayes Estimates of Cross-center Parameter Distributions



A. First stage



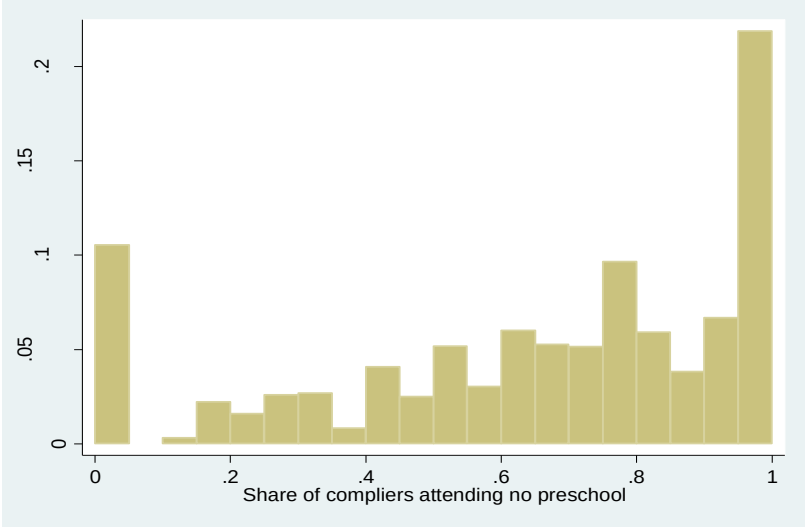
B. Cognitive effect



C. Non-cognitive effect

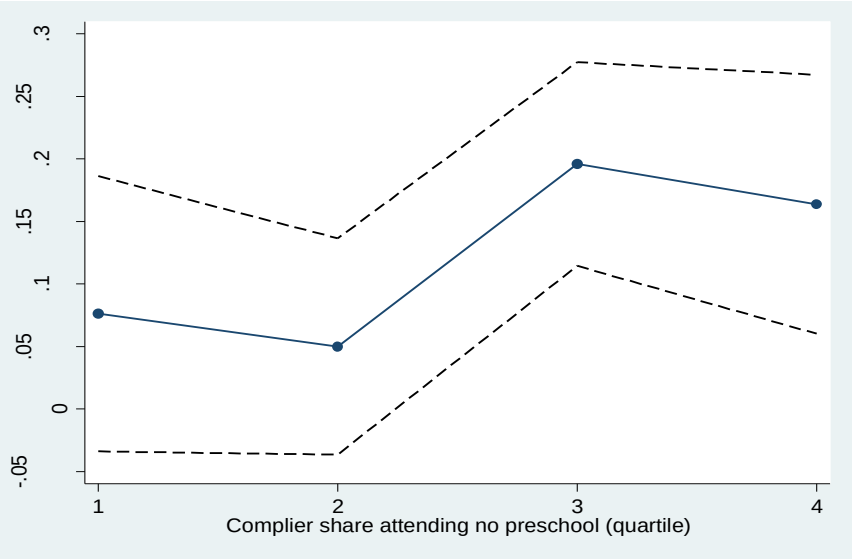
Notes: This figure plots Empirical Bayes maximum simulated likelihood estimates of the cross-center distributions of parameters in Spring 2003. Red curves are kernel density estimates produced using 10,000 draws from the prior distributions listed in Table A2.

Figure 5: Histogram of Share of Compliers Attending No Preschool

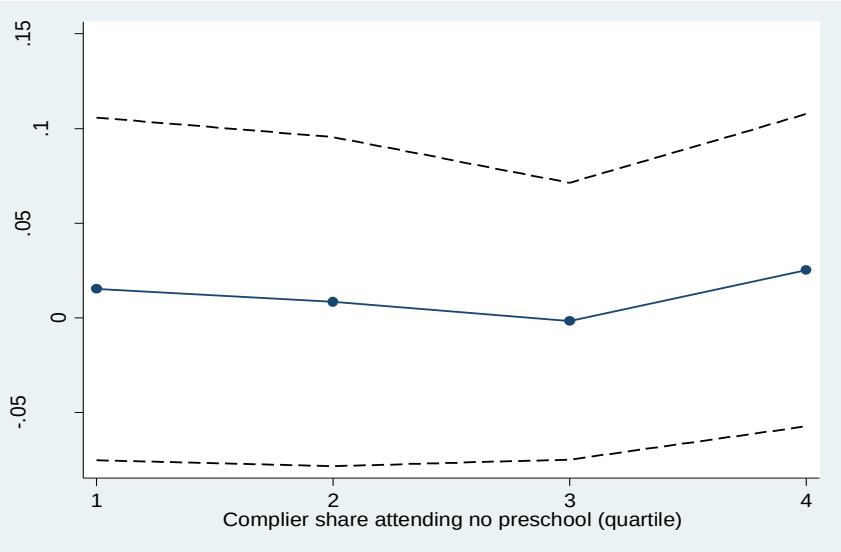


Notes: This figure shows the histogram of center-specific shares of compliers attending no preschool. The data are grouped into bins of width 0.05.

Figure 6: Relationships Between Head Start Effects and Complier Counterfactuals (Spring 2003)



A. Cognitive effects



B. Non-cognitive effects

Notes: This figure plots coefficients from two-stage least squares models that interact indicators for quartiles of the center-specific share of compliers attending no preschool with Head Start attendance. The instruments are interactions of quartiles and a Head Start offer indicator. Dashed lines plot 95% confidence intervals. The $\chi^2(3)$ statistics from tests of the hypothesis that effects are equal across quartiles are 7.00 (p = 0.072) and 0.24 (p = 0.97) for cognitive and non-cognitive skills.

Table 1: Outcomes Included in Summary Indices

Cognitive skills	Non-cognitive skills
(1)	(2)
Peabody Picture and Vocabulary Test III (PPVT)	Takes care of personal things
Color names	Asks for assistance with tasks
Test de Vocabularioen Imagenes Peabody (TVIP adapted)	Makes friends easily
Woodcock-Johnson III Oral Comprehension	Enjoys learning
Preschool Comprehensive Test of Phonological and Print Processing (CTOPPP)	Has temper tantrums
Spanish CTOPPP	Cannot concentrate/pay attention for long
Woodcock-Johnson III Word Attack	Is very restless/fidgets a lot
McCarthy Draw-a-design	Likes to try new things
Letter naming	Shows imagination in work and play
Woodcock-Johnson III Letter-Word Identification	Hits and fights with others
Bateria R Woodock-Munoz Identificacion de Letras y Palabras	Accepts friends' ideas in playing
Woodcock-Johnson III Spelling	
Bateria R Woodcock-Munoz Dictado	
Woodcock-Johnson III Applied Problems	
Woodcock-Johnson III Quantitative Concepts	
Counting Bears	
Bateria R Woodcock-Munoz Problemas Aplicados	

Notes: This table lists the cognitive and non-cognitive outcomes used in the analysis. Summary indices are averages of standardized outcomes in each category.

Table 2: Characteristics of Head Start Applicants

Variable	Control mean (1)	Offer differential (2)
Male	0.490	0.011
		(0.022)
Black	0.259	0.009
		(0.018)
Hispanic	0.411	0.000
		(0.022)
Special needs	0.112	0.020
		(0.014)
Mother is married	0.478	-0.016
		(0.022)
Both parents live at home	0.531	-0.016
		(0.021)
Teen mother	0.165	-0.023
		(0.015)
Mother is high school dropout	0.389	-0.022
		(0.021)
Mother attended college	0.281	0.020
		(0.019)
Monthly household income	1576.370	-16.343
		(72.850)
Baseline cognitive skills	-0.003	0.014
		(0.024)
Baseline non-cognitive skills	0.001	0.028
		(0.017)
Attended Head Start in 1st year	0.160	0.663***
		(0.017)
Attended any preschool in 1st year	0.460	0.442***
		(0.020)
Joint *p* -value for baseline characteristics	-	0.508
N (total)	4,442	
N (completed survey)	3,577	

Notes: Column (1) shows means of baseline characteristics for Head Start applicants assigned to the control group. Column (2) shows coefficients from regressions of each characteristics on assignment to Head Start. The means and regressions are weighted using the HSIS baseline child weights. The *p* -value is from a test of the hypothesis that coefficients for all baseline characteristics are zero.

***significant at 1%; **significant at 5%; *significant at 10%

Table 3: Characteristics of Head Start Centers

	Head Start centers				Private centers
	Mean	Std. dev.	Min.	Max.	Mean
Variable	(1)	(2)	(3)	(4)	(5)
Fraction of teachers with bachelor's degree	0.35	0.40	0.00	1.00	0.41
Fraction of staff with teaching license	0.11	0.23	0.00	1.00	0.30
Student/staff ratio	6.79	1.71	2.33	13.50	8.76
Full day service	0.63	0.48	0.00	1.00	0.67
More than three home visits per year	0.20	0.40	0.00	1.00	0.13
High/Scope curriculum	0.30	0.46	0.00	1.00	0.28
Number of randomized applicants	12.90	10.46	2.00	79.00	-
Fraction of applicants assigned to Head Start	0.59	0.06	0.25	0.83	-
N (centers)	302				

Notes: This table summarizes characteristics of Head Start center in the HSIS data. Means and standard deviations are student-weighted for variables other than number of applicants and fraction assigned to Head Start. The HSIS sample excludes centers where the center director did not answer the HSIS survey, and centers where the fraction of students assigned to Head Start was zero or one.

Table 4: Average Effects of Head Start on Cognitive and Non-cognitive Skills by Cohort and Year

		Cognitive skills		Non-cognitive skills	
		First stage	IV estimate	First stage	IV estimate
Time period	Cohort	(1)	(2)	(3)	(4)
Spring 2003	3-year-olds	0.679***	0.172***	0.680***	0.052*
		(0.025)	(0.038)	(0.025)	(0.031)
		2070	2070	2062	2062
	4-year-olds	0.684***	0.090**	0.685***	-0.041
		(0.027)	(0.038)	(0.027)	(0.036)
		1638	1638	1631	1631
Spring 2004	3-year-olds	0.362***	0.157**	0.359***	0.078
		(0.026)	(0.078)	(0.026)	(0.070)
		2046	2046	2032	2032
	4-year-olds	0.692***	-0.079	0.692***	-0.035
		(0.028)	(0.049)	(0.027)	(0.042)
		1535	1535	1555	1555
Spring 2005	3-year-olds	0.376***	-0.008	0.380***	0.043
		(0.027)	(0.098)	(0.027)	(0.074)
		1927	1927	1996	1996
	4-year-olds	0.668***	0.006	0.668***	-0.065
		(0.028)	(0.059)	(0.028)	(0.042)
		1527	1527	1576	1576
Spring 2006	3-year-olds	0.368***	0.064	0.372***	0.031
		(0.027)	(0.111)	(0.027)	(0.071)
		1876	1876	1957	1957

Notes: This table reports estimates of the effect of Head Start attendance on summary indices of cognitive and non-cognitive skills. Estimates come from instrumental variables models using assignment to Head Start as an instrument for Head Start attendance. All models use the HSIS baseline child weights. Robust standard errors in parentheses.
***significant at 1%; **significant at 5%; *significant at 10%

Table 5: Finite-sample Behavior of Center-specific Instrumental Variables Estimates

	Mean	Std. dev.	Min.	Max.
	(1)	(2)	(3)	(4)
IV estimate	0.211	1.390	-4.585	15.236
IV asymptotic standard error	1.246	6.204	0.038	92.968
Implied cross-center variance of effects		-18.275		

Notes: This table summarizes the distribution of center-specific instrumental variables estimates for cognitive skills in Spring 2003. The estimate for each center is a separate IV regression of cognitive skills on Head Start attendance instrumented by assignment, pooling the 3- and 4-year-old cohorts and using the HSIS child weights. The implied cross-center variance of effects is the sample variance of the IV estimates minus the average squared standard error. The sample excludes centers with less than 3 applicants and centers with first stages equal to exactly zero. Two other centers with small samples and first stages very close to zero are also dropped. The sample includes 285 centers.

Table 6: Comparison of Two-stage Least Squares and Maximum Likelihood Estimates

	Two-stage least squares		Maximum likelihood		
	First stage	LATE	First stage	LATE	ATE
Outcome (Spring 2003)	(1)	(2)	(3)	(4)	(5)
Cognitive skills	0.719***	0.121***	0.719***	0.119***	0.119***
	(0.012)	(0.031)	(0.012)	(0.030)	(0.030)
		3708		3708	
Non-cognitive skills	0.719***	0.026	0.719***	0.026	0.026
	(0.012)	(0.020)	(0.012)	(0.020)	(0.020)
		3693		3693	

the selection model described in the text with no cross-center heterogeneity. The sample pools three- and four-year-old cohorts, and all models use the HSIS baseline child weights. Robust standard errors in parentheses.
***significant at 1%; **significant at 5%; *significant at 10%

Table 7: Random Coefficients Estimates for Spring 2003

		Cognitive skills		Non-cognitive skills	
		Estimate	Standard error	Estimate	Standard error
Parameter	Description	(1)	(2)	(3)	(4)
$E[\Phi(\lambda_j+\pi_j) - \Phi(\lambda_j)]$	Mean compliance probability	0.750***	0.020	0.742***	0.020
$[Var(\Phi(\lambda_j+\pi_j) - \Phi(\lambda_j))]^{1/2}$	Std. dev. of compliance probability	0.205***	0.011	0.206***	0.011
$E[\alpha_{1j}]$	Mean treated outcome	0.112***	0.023	0.024	0.016
$E[\alpha_{0j}]$	Mean non-treated outcome	-0.011	0.029	-0.002	0.016
$E[\alpha_{1j} - \alpha_{0j}]$	Mean Head Start effect	0.123***	0.033	0.026	0.020
$[Var(\alpha_{1j} - \alpha_{0j})]^{1/2}$	Std. dev. of Head Start effects	0.181***	0.014	0.066***	0.007

Notes: This table lists maximum simulated likelihood estimates of parameters of the cross-center distribution of Head Start effects in Spring 2003. The sample pools the three- and four-year-old cohorts, and observations are weighted using the HSIS baseline child weights. The MSL procedure uses 1,000 simulations for each Head Start center. Robust standard errors in parentheses.
***significant at 1%; **significant at 5%; *significant at 10%

Table 8: Relationships Between Inputs and Head Start Effects

	Cognitive skills			Non-cognitive skills		
	Two-stage least squares		Maximum	Two-stage least squares		Maximum
	Bivariate	Multivariate	likelihood	Bivariate	Multivariate	likelihood
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Fraction of staff with bachelor's degree	-0.012	0.021	0.086	-0.026	-0.047	-0.052
	(0.076)	(0.072)	(0.063)	(0.058)	(0.057)	(0.040)
Fraction of staff with teaching license	-0.151	-0.089	-0.113	0.124	0.115	0.083
	(0.117)	(0.115)	(0.105)	(0.089)	(0.091)	(0.066)
Student/staff ratio	-0.013	-0.003	0.023	0.000	0.000	0.012
	(0.016)	(0.016)	(0.015)	(0.012)	(0.012)	(0.017)
Full day service	0.137**	0.133**	0.130**	-0.042	-0.058	-0.009
	(0.058)	(0.060)	(0.053)	(0.048)	(0.049)	(0.033)
More than three home visits per year	0.025	0.026	0.080	0.118**	0.134***	0.088**
	(0.062)	(0.063)	(0.063)	(0.049)	(0.051)	(0.040)
High/Scope curriculum	-0.007	-0.022	-0.044	0.042	0.066	0.016
	(0.061)	(0.061)	(0.055)	(0.048)	(0.050)	(0.034)
Residual std. dev. of Head Start effects	-	-	0.154	-	-	0.056
R-squared			0.275			0.285

Notes: This table reports estimates of relationships between Head Start effects and inputs in Spring 2003. Two-stage least squares models instrument Head Start attendance and its interactions with inputs using assignment to Head Start and its interactions with inputs. Columns (1) and (4) estimate a separate interaction model for each input, while columns (2) and (5) include all interactions simultaneously. Main effects of interacting variables are included as controls. All models weight observations using the HSIS baseline child weights. Robust standard errors in parentheses.

***significant at 1%; **significant at 5%; *significant at 10%

Table 9: Relationship Between Head Start Effects and Complier Counterfactuals

	Share of compliers attending no preschool:	
	Below median	Above median
Outcome (Spring 2003)	(1)	(2)
Cognitive skills	0.062*	0.177***
	(0.035)	(0.035)
N	3424	
p -value	0.020	
Non-cognitive skills	0.029	0.021
	(0.032)	(0.029)
N	3418	
p -value	0.760	

Notes: This table shows two-stage least squares estimates of models interacting Head Start attendance with an indicator equal to one if the share of compliers drawn from attending no preschool at a child's center is above the sample median. The fractions of children who are no-preschool and private-preschool compliers are estimated by regressing dummies for no preschool and private preschool on a Head Start offer dummy, respectively; the complier no-preschool share is the ratio of the coefficient from the no-preschool regression to the sum of the two coefficients (with coefficients above zero set to zero). The instruments are the Head Start offer and the interaction of the offer with an above-median indicator. All models use the HSIS baseline child weights. Robust standard errors in parentheses.

***significant at 1%; **significant at 5%; *significant at 10%

References

1. Anderson, M. (2008). "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103(484).

2. Angrist, J., and Imbens, G. (1995). "Two-stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90(430).

3. Angrist, J., and Lavy, V. (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114(2).

4. Berruta-Clement, J., Schweinhart, L., Barnett, W., Epstein, A., and Weikart, D. (1984). Changed Lives: The Effects of the Perry Preschool Program on Youths Through Age 19. Ypsilanti, MI: High/Scope Press.

5. Bertrand, M., and Pan, J. (2013). "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior." *American Economic Journal: Applied Economics* 5(1).

6. Bitler, M., Domina, T., and Hoynes, H. (2012). "Experimental Evidence on Distributional Effects of Head Start." Mimeo, University of California, Irvine.

7. Campbell, F., and Ramey, C. (1994). "Effects of Early Intervention on Intellectual and Academic Achievement: A Follow-up Study of Children from Low-Income Families." *Child Development* 65(2).

8. Campbell, F., and Ramey, C. (1995). "Cognitive and School Outcomes for High-Risk African-American Students at Middle Adolescence: Positive Effects of Early Intervention." *American Educational Research Journal* 32(4).

9. Cascio, E., and Schanzenbach, D. (2013). "The Impacts of Expanding Access to High-Quality Preschool Education." Brookings Papers on Economic Activity.

10. Chetty, R., Hilger, N., Saez, E., Schanzenbach, D., and Yagan, D. (2011). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4).

11. Chetty, R., Friedman, J., and Rockoff, J. (2013a). "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-added Estimates." NBER Working Paper no. 19423.

12. Chetty, R., Friedman, J., and Rockoff, J. (2013b). "Measuring the Impacts of Teachers II: Teacher Value-added and Student Outcomes in Adulthood." NBER Working Paper no. 19424.

13. Currie, J., and Thomas, D. (1995). "Does Head Start Make a Difference?" *American Economic Review* 85(3).

14. Deming, D. (2009). "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3).

15. Deming, D. (2013). "Using School Choice Lotteries to Test Measures of School Effectiveness." National Bureau of Economic Research Working Paper no. 19803.

16. Epstein, A. (2007). Essentials of Active Learning in Preschool: Getting to Know the High/Scope Curriculum. Ypsilanti, MI: High/Scope Press.

17. Fitzpatrick, M. (2008). "Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement." *The B.E. Journal of Economic Analysis and Policy* 8(1).

18. Garces, E., Thomas, D., and Currie, J. (2002). "Longer-term Effects of Head Start." *American Economic Review* 92(4).

19. Gelber, A., and Isen, A. (2013). "Children's Schooling and Parents'Behavior: Evidence from the Head Start Impact Study." *Journal of Public Economics* 101.

20. Geweke, J., Gowrisankaran, G., and Town, R. (2003). "Bayesian Inference for Hospital Quality in a Selection Model." *Econometrica* 71(4).

21. Gibbs, C., Ludwig, J., and Miller, D. (2011). "Does Head Start Do Any Lasting Good?" National Bureau of Economic Research Working Paper no. 17452.

22. Gormley, W., and Gayer, T. (2005). "Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program." *Journal of Human Resources* 40.

23. Hanushek, E. (2009). "Teacher Deselection." In: Creating a New Teaching Profession, D. Goldhaber and J. Hannaway, eds. Washington, DC: Urban Institute Press.

24. Heckman, J. (1979). "Sample Selection Bias as a Specification Error." *Econometrica* 47(1).

25. Heckman, J. (2011). "The American Family in Black and White: A Post-Racial Strategy for Improving Skills to Promote Equality." IZA Discussion Paper no. 5495.

26. Heckman, J., Moon, S., Pinto, R., Savelyev, P., and Yavitz, A. (2010a). "The Rate of Return to the High/Scope Perry Preschool Program." *Journal of Public Economics* 94.

27. Heckman, J., Moon, S., Pinto, R., Savelyev, P., and Yavitz, A. (2010b). "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the High/Scope Perry Preschool Program." *Quantitative Economics* 1(1).

28. Heckman, J., Malofeeva, L., Pinto, R., and Savelyev, P. (2013). "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103(6).

29. Heckman, J., Urzua, S., and Vytlacil, E. (2006). "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics* 88(3).

30. Imbens, G., and Angrist, J., (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2).

31. Imbens, G., and Rubin, D. (1997). "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *Annals of Statistics* 25(1).

32. Jacob, B. (2002). "Where the Boys Aren't: Non-cognitive Skills, Returns to School and the Gender Gap in Higher Education." *Economics of Education Review* 21.

33. Jacob, B., and Lefgren, L. (2008). "Principals as Agents: Subjective Performance Assessment in Education." *Journal of Labor Economics* 26(1).

34. Kane, T., Rockoff, J., and Staiger, D. (2008). "What does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27(6).

35. Kling, J., Liebman, J., and Katz, L. (2007). "Experimental Analysis of Neighborhood Effects." *Econometrica* 75(1).

36. Krueger, A. (1999). "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114(2).

37. Ludwig, J., and Miller, D. (2007). "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *Quarterly Journal of Economics* 122(1).

38. Mariano, R. (1977). "Finite Sample Properties of Instrumental Variable Estimators of Structural Coefficients." *Econometrica* 45(2).

39. Morris, C. (1983). "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78(381).

40. Nelson, C., and Startz, R. (1990). "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58(4).

41. Obama, B. (2013). The White House, Office of the Press Secretary. Remarks by the President in state of the union address.

42. Raudenbush, S., Reardon, S., and Nomi, T. (2012). "Statistical Analysis for Multisite Trials Using Instrumental Variables with Random Coefficients." *Journal of Research on Educational Effectiveness* 5(3).

43. Reynolds, A. (1998). "Extended Early Childhood Intervention and School Achievement: Age Thirteen Findings from the Chicago Longitudinal Study." *Child Development* 69(1).

44. Schweinhart, L. (2007). "How to Take the High/Scope Perry Preschool to Scale." Paper prepared for the National Invitational Conference of the Early Childhood Research Collaborative.

45. Schweinhart, L., Montie, J., Xiang, Z., Barnett, W., Belfield, C., and Nores, M. (2005). Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40. Ypsilanti, MI: High/Scope Press.

46. Schweinhart, L., and Weikart, D. (1997). Lasting Differences: The High/Scope Preschool Curriculum Comparison Study Through Age 23. Ypsilanti, MI: High/Scope Press.

47. Train, K. (2003). Discrete Choice Models with Simulation. New York: Cambridge University Press.

48. US Department of Health and Human Services, Administration for Children and Families (2008). "Statutory Degree and Credentialing Requirements for Head Start Teaching Staff." http://eclkc.ohs.acf.hhs.gov/hslc/standards/IMs_and_PIs_in_PDF/PDF_IMs/IM2008/ACF-IM-HS-08-12.pdf. Accessed March 27, 2013.

49. US Department of Health and Human Services, Administration for Children and Families (2010). "Head Start Impact Study, Final Report." Washington, DC.

50. US Department of Health and Human Services, Administration for Children and Families (2011). "Head Start Program Facts, Fiscal Year 2011." http://eclkc.ohs.acf.hhs.gov/hslc/mr/factsheets/docs/hs-program-fact-sheet-2011-final.pdf. Accessed March 27, 2013.

51. US Department of Health and Human Services, Administration for Children and Families (2012). "Third Grade Follow-up to the Head Start Impact Study." Washington, DC .

52. US Office of Head Start (2012). "Head Start Services." http://www.acf.hhs.gov/programs/ohs/about/head-start. Accessed March 27, 2013.

53. Wong, V., Cook, T., Barnett, W., and Jung, K. (2008). "An Effectiveness-based Evaluation of Five State Pre-kindergarten Programs." *Journal of Policy Analysis and Management* 27(1).

Table A1: Attrition by Cohort and Year

		Cognitive skills		Non-cognitive skills	
		Follow-up rate	Differential	Follow-up rate	Differential
Time period	Cohort	(1)	(2)	(3)	(4)
Spring 2003	3-year-olds	0.845	0.029*	0.842	0.002
			(0.015)		(0.015)
			2449		2449
	4-year-olds	0.822	0.030*	0.818	-0.002
			(0.018)		(0.017)
			1993		1993
Spring 2004	3-year-olds	0.835	0.021	0.830	0.024
			(0.019)		(0.019)
			2449		2449
	4-year-olds	0.770	0.009	0.780	-0.001
			(0.022)		(0.021)
			1993		1993
Spring 2005	3-year-olds	0.787	-0.003	0.815	0.006
			(0.019)		(0.018)
			2449		2449
	4-year-olds	0.766	0.010	0.791	0.012
			(0.025)		(0.024)
			1993		1993
Spring 2006	3-year-olds	0.766	0.016	0.799	0.030
			(0.020)		(0.020)
			2449		2449

Notes: This table reports attrition rates for the HSIS sample. Columns (1) and (3) show fractions of children with observed outcomes by cohort and time period. Columns (2) and (4) report treatment/control differences. These differences are coefficients on a treatment indicator from regressions of a dummy for an observed outcome on treatment status, with the same controls and weighting scheme as in Table 4.

***significant at 1%; **significant at 5%; *significant at 10%

Table A2: Estimates of Pooled Selection Model for Spring 2003

		Cognitive skills		Non-cognitive skills	
		Estimate	Standard error	Estimate	Standard error
Parameter	Description	(1)	(2)	(3)	(4)
α_1	Mean treated outcome	0.102***	0.017	0.025*	0.013
α_0	Mean control outcome	-0.017	0.024	-0.001	0.015
λ	Intercept in selection equation	-1.055***	0.041	-1.072***	0.041
π	Offer coefficient in selection equation	2.155***	0.053	2.157***	0.053
σ_1	Std. dev. of treated outcome	0.598***	0.009	0.426***	0.007
σ_0	Std. dev. of control outcome	0.656***	0.012	0.424***	0.008
ρ_1	Correlation between treated outcome and selection error	0.126**	0.052	0.022	0.057
ρ_0	Correlation between control outcome and selection error	0.120**	0.047	-0.001	0.046

Notes: This table lists maximum likelihood estimates of the selection model described in the text. The sample pools the three- and four-year-old cohorts, and observations are weighted using the HSIS baseline child weights. Robust standard errors in parentheses.

Table A3: Empirical Bayes Estimates By Time Period

		Cognitive skills		Non-cognitive skills	
		Spring 2003	Spring 2005	Spring 2003	Spring 2005
Parameter	Description	(1)	(2)	(3)	(4)
$E[\alpha_{1j}]$	Mean treated outcome	0.112***	-0.011	0.024	0.017
		(0.023)	(0.027)	(0.016)	(0.017)
$E[\alpha_{0j}]$	Mean non-treated outcome	-0.011	-0.025	-0.002	0.009
		(0.023)	(0.027)	(0.016)	(0.017)
$E[\lambda_j]$	Mean of intercept in selection equation	-1.355***	-0.413***	-1.342***	-0.429***
		(0.023)	(0.027)	(0.016)	(0.017)
$E[\log\pi_j]$	Mean of log of offer coefficient in selection equation	0.850***	0.400***	0.833***	0.412***
		(0.023)	(0.027)	(0.016)	(0.017)
$[Var(\alpha_{1j})]^{1/2}$	Std. dev. of mean treated outcome	0.220***	0.249***	0.106***	0.080***
		(0.019)	(0.022)	(0.014)	(0.016)
$[Var(\alpha_{0j})]^{1/2}$	Std. dev. of mean non-treated outcome	0.258***	0.264***	0.091***	0.074***
		(0.024)	(0.033)	(0.018)	(0.023)
$[Var(\lambda_j)]^{1/2}$	Std. dev. of intercept in selection equation	0.865***	0.504***	0.873***	0.482***
		(0.098)	(0.102)	(0.098)	(0.097)
$[Var(\log\pi_j)]^{1/2}$	Std. dev. of log of offer coefficient in selection equation	0.509***	0.435***	0.511***	0.403***
		(0.051)	(0.067)	(0.052)	(0.065)
σ_1	Std. dev. of error in treated equation	0.579***	0.650***	0.413***	0.451***
		(0.009)	(0.010)	(0.007)	(0.007)
σ_0	Std. dev. of error in non-treated equation	0.625***	0.701***	0.415***	0.461***
		(0.013)	(0.016)	(0.008)	(0.010)
ρ_1	Correlation between treated outcome and selection error	0.108	-0.033	0.023	0.043
		(0.076)	(0.072)	(0.080)	(0.075)
ρ_0	Correlation between control outcome and selection error	0.150***	0.077	-0.008	0.043
		(0.053)	(0.071)	(0.050)	(0.066)

This table lists Empirical Bayes maximum simulated likelihood estimates of parameters of the cross-center distribution of Head Start effects by year. The sample pools the three- and four-year-old cohorts, and observations are weighted using the HSIS baseline child weights. The MSL procedure uses 1,000 simulations for each Head Start center. Robust standard errors in parentheses.

***significant at 1%; **significant at 5%; *significant at 10%

Table A4: Empirical Bayes Estimates for Spring 2005

		Cognitive skills		Non-cognitive skills	
		Estimate	Standard error	Estimate	Standard error
Parameter	Description	(1)	(2)	(3)	(4)
$E[\Phi(\lambda_j+\pi_j) - \Phi(\lambda_j)]$	Mean compliance probability	0.520***	0.022	0.526***	0.021
$[Var(\Phi(\lambda_j+\pi_j) - \Phi(\lambda_j))]^{1/2}$	Std. dev. of compliance probability	0.188***	0.011	0.177***	0.011
$E[\alpha_{1j}]$	Mean treated outcome	-0.011	0.027	0.017	0.017
$E[\alpha_{0j}]$	Mean non-treated outcome	-0.025	0.049	0.009	0.029
$E[\alpha_{1j} - \alpha_{0j}]$	Mean Head Start effect	0.014	0.052	0.007	0.032
$[Var(\alpha_{1j} - \alpha_{0j})]^{1/2}$	Std. dev. of Head Start effects	0.065**	0.033	0.033***	0.007

Notes: This table lists Empirical Bayes maximum simulated likelihood estimates of parameters of the cross-center distribution of Head Start effects in Spring 2003. The sample pools the three- and four-year-old cohorts, and observations are weighted using the HSIS baseline child weights. The MSL procedure uses 1,000 simulations for each Head Start center. Robust standard errors in parentheses.
***significant at 1%; **significant at 5%; *significant at 10%

Table A5: Maximum Likelihood Estimates of Finite-type Models

		Three-type model			Five-type model				
		Type 1	Type 2	Type 3	Type 1	Type 2	Type 3	Type 4	Type 5
Parameter	Description	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
α_1^k	Mean treated outcome	-0.055**	0.385***	0.069	-0.059**	0.355***	0.408***	0.116	0.106
		(0.027)	(0.033)	(0.054)	(0.027)	(0.039)	(0.095)	(0.104)	(0.064)
α_0^k	Mean control outcome	-0.184***	0.249***	0.281**	-0.171***	0.318***	-0.144	0.925***	0.005
		(0.032)	(0.038)	(0.133)	(0.032)	(0.043)	(0.109)	(0.192)	(0.139)
α_1^k - α_0^k	Head Start effect	0.130***	0.135***	-0.211	0.112***	0.037	0.552***	-0.809***	0.101
		(0.039)	(0.046)	(0.131)	(0.040)	(0.052)	(0.137)	(0.199)	(0.154)
$\Phi(\lambda^k+\pi^k)$ - $\Phi(\lambda^k)$	Compliance probability	0.760***	0.865***	0.214***	0.748***	0.842***	0.936***	0.552***	0.012
		(0.020)	(0.019)	(0.079)	(0.022)	(0.024)	(0.033)	(0.103)	(0.099)
P^k	Type probability	0.524***	0.344***	0.133***	0.518***	0.263***	0.097**	0.034**	0.088***
		(0.051)	(0.045)	(0.032)	(0.052)	(0.047)	(0.039)	(0.017)	(0.027)
$\left[\sum P^k\left((\alpha_1^k-\alpha_0^k)-(\overline{\alpha_1}-\overline{\alpha_0})\right)^2\right]^{.5}$	Std. dev. of Head Start effects		0.116				0.222		

Notes: This table reports maximum likelihood estimates of finite-type models for cognitive skills in Spring 2003. Columns (1)-(3) come from a model assuming Head Start centers belong to one of three types, while columns (4)-(8) come from a model assuming center belong to one of five types. Robust standard errors in parentheses.

Online Appendix

The data for this analysis come from the Head Start Impact Study (HSIS). The HSIS data includes information on 4,442 students. Each student applied to one of 353 Head Start centers in Fall 2002, and each center is associated with one of 84 regional Head Start program areas. The data includes separate files with information on test scores, answers to parental surveys, and Head Start center characteristics. This Appendix describes the procedure used to clean each data source and construct the data set used for analysis.

Test Score Data

Test score information comes from a series of assessments conducted in Fall 2002, Spring 2003, Spring 2004, Spring 2005 and Spring 2006. From each assessment file, I extract raw scores for the 17 tests listed in column (1) of Table 1. These 17 tests are the main outcomes examined by DHHS (2010). The data also include a few other tests (for example, the Leiter Sustained Attention Task), but DHHS (2010) expresses reservations about their reliability and hence they are excluded. Not all tests were administered every year, and there were some differences in the tests administered to Spanish-speaking and English-speaking students; for example, the TVIP and Spanish CTOPPP were administered to Spanish speakers only. To construct the cognitive summary index outcome, I standardize each test relative to the control group among students who took the test separately for each cohort and assessment period. I then compute the mean of observed standardized outcomes for each child. Finally, I append together the data sets for each assessment period, and use a unique student identifier to reshape the data into a wide format file with one observation per student and a separate variable for the cognitive summary index in each assessment period.

Parent Survey Data

Baseline demographics

Information on student demographics is drawn from a baseline survey of parents conducted in Fall 2002. Eighty-one percent of households responded to this survey (3,577 of 4,442). This demographic information is supplemented with a set of derived variables from the HSIS "Covariates and Subgroups" data file. This file combines the baseline survey with information collected during experimental recruitment to fill in missing values for some demographic variables. When variables are present in both files, information from the "Covariates and Subgroups" file is used.

Non-cognitive outcomes

Indices of non-cognitive skill are constructed from the baseline parental survey and follow-up surveys conducted in Spring 2003, Spring 2004, Spring 2005 and Spring 2006. I begin with the all social and emotional outcomes analyzed by DHHS (2010). Each outcome is redefined so that a positive sign is favorable, and then standardized relative to the control group separately by cohort and survey period. I also retain raw measures of each outcome. I then append together the files for all periods. To exclude outcomes without

meaningful variation, I compute the mean of each raw outcome over all survey periods, and drop outcomes where more than 90% of responses were the same. This produces the set of outcomes listed in column (2) of Table 1. I then compute the non-cognitive summary index for each survey period as the mean of the remaining standardized outcomes. Finally, I use the unique student identifier to reshape the data into a wide format file with one observation per student and a separate variable for the non-cognitive summary index in each survey period.

Measuring Head Start Assignment and Attendance

Head Start assignment comes from an administrative variable generated at the time of random assignment. Head Start attendance in Spring 2003 is also measured administratively. To measure Head Start attendance in later periods, I combine this administrative measure with parental survey information. Specifically, I set Head Start attendance equal to one for Spring 2004, Spring 2005 and Spring 2006 if the Spring 2003 administrative measure is one, or if a parent indicated Head Start attendance at any time up to the relevant time period. For time periods after Spring 2003, the Head Start attendance variable is missing for children whose parents did not respond to the survey, because attendance cannot be accurately measured for these students. This restriction does not affect the main results, which focus on Spring 2003.

Center Characteristics

The characteristics of Head Start centers are measured from a childcare center director survey conducted in Spring 2003. The survey attempted to collect information from directors of all childcare centers attended by sample children, including members of the control group who attended childcare outside of Head Start centers in the experimental sample. The director survey data set is a student-level file, with variables capturing responses of the center director at the center attended by each child. The six key inputs listed in Table 3 are derived from the following questions:

- **High/Scope curriculum:** "If your principal curriculum has a name, what is that name?" Centers are coded as High/Scope if the director selected High/Scope from among a list of possible answers to this question.
- **Fraction of staff with bachelors degree:** "Approximately what percentage of lead and assistant teachers in your center have a bachelors degree or higher?"
- **Fraction of staff with teaching license:** "Approximately what percentage of lead and assistant teachers in your center have a teaching certificate or license?"
- **Student/staff ratio:** This variable is derived from answers to four questions. I measure center capacity using the answer to the question: "What is the center's preschool service capacity?" While capacity may not always equal enrollment, 94 percent of responses indicated that the center was filled

to capacity "all the time" or "most of the time." Staff size is contructed by adding together answers tothree questions of the form: "How many X are currently employed at the center?" where X is "lead teachers," "assistant teachers," or "paid teacher aides." The student/staff ratio is then constructed by dividing capacity by staff size.

- **Full-day service:** "What child care options are provided at the center?" Centers are coded as full-day if the director selected "full-day" from a list of possible responses to this question.
- **More than three home visits per year:** "How many home visits are required per program year?" Directors were given a list of possible responses to this question. About 1 percent of responses were "1 visit," 79 percent of responses were "2-3 visits," and 20 percent of responses were "more than three visits."

I use these questions to derive the characteristics of each center of random assignment. To this end, I keep observations administratively coded as both assigned to the treatment group and attending Head Start. In some cases, codes for the center director were different for such students within a center of random assignment. I use responses for the center director most frequently associated with treated students at a given center of random assignment. For 7 percent of centers, there were two center director interviews associated with an equal number of treated students. I break ties randomly to determine which responses to use in these cases. I then keep one observation per center of random assignment. The resulting data set has information for 89 percent (314 out of 353) of centers in the HSIS experiment.

Constructing the Analysis Data Set

The procedure described above yields 5 data files: A test score file, a baseline demographic file, a non-cognitive outcome file, a file coding Head Start attendance after Spring 2003, and a center characteristics file. I merge the first four of these files using a unique student identifier. I then merge the resulting file with the center characteristics file using an identifier for center of random assignment. Finally, I merge on a sixth file containing the HSIS baseline child weights, which yields the final data set used for analysis.